Comparison of Targeted Maximum Likelihood and Shrinkage Estimators of Parameters in Gene Networks

Geeven, G.; van der Laan, M.J.; de Gunst, M.C.M.

published in
Statistical Applications in Genetics and Molecular Biology
2012

DOI (link to publisher)
10.1515/1544-6115.1728

document version
Publisher's PDF, also known as Version of record

Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 5 2012 *Article* 2

Comparison of Targeted Maximum Likelihood and Shrinkage Estimators of Parameters in Gene Networks

Geert Geeven, *VU University Amsterdam*
Mark J. van der Laan, *University of California - Berkeley*
Mathisca C.M. de Gunst, *VU University Amsterdam*

Recommended Citation:
Geeven, Geert; van der Laan, Mark J.; and de Gunst, Mathisca C.M. (2012) "Comparison of Targeted Maximum Likelihood and Shrinkage Estimators of Parameters in Gene Networks," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 5, Article 2.
DOI: 10.1515/1544-6115.1728

Comparison of Targeted Maximum Likelihood and Shrinkage Estimators of Parameters in Gene Networks

Geert Geeven, Mark J. van der Laan, and Mathisca C.M. de Gunst

Abstract

Gene regulatory networks, in which edges between nodes describe interactions between transcription factors (TFs) and their target genes, model regulatory interactions that determine the cell-type and condition-specific expression of genes. Regression methods can be used to identify TF-target gene interactions from gene expression and DNA sequence data. The response variable, i.e. observed gene expression, is modeled as a function of many predictor variables simultaneously. In practice, it is generally not possible to select a single model that clearly achieves the best fit to the observed experimental data and the selected models typically contain overlapping sets of predictor variables. Moreover, parameters that represent the marginal effect of the individual predictors are not always present. In this paper, we use the statistical framework of estimation of variable importance to define variable importance as a parameter of interest and study two different estimators of this parameter in the context of gene regulatory networks. On yeast data we show that the resulting parameter has a biologically appealing interpretation. We apply the proposed methodology on mammalian gene expression data to gain insight into the temporal activity of TFs that underly gene expression changes in F11 cells in response to Forskolin stimulation.

KEYWORDS: gene regulatory networks, variable importance, targeted maximum likelihood

Author Notes: This work received financial support from the Netherlands Organization for Scientific Research (NWO; CLS grant 635.100.008 to MCMdG), from the NCA (Neuroscience Campus Amsterdam) and from NIAID (grant R01AI74345-5 to MvdL).

1 Introduction

Cell-type- and condition-specific interactions between transcriptional regulators and their target genes are a primary mechanism for cells to accomplish spatiotemporal changes in gene expression. Regression models, in which predictor variables represent *in silico* predicted transcription factor (TF) binding affinity, can be used to study the effect of TF binding on observed gene expression of target genes (Bussemaker, Li, and Siggia, 2001, Das, Banerjee, and Zhang, 2004). Such models describe gene expression as a function of many predictors simultaneously and are typically used to answer two important questions, i.e. *which* TFs are associated to a given gene expression response of interest and, secondly, which of them are *most important*. In practice, there are usually several candidate models that fit almost equally well and that contain different, partially overlapping, sets of predictors. Moreover, the predictors typically occur in many model terms and a single term that can be interpreted as a marginal effect, such as a main effect term, is often lacking. Therefore, from these fitted models, it is not clear how to rank candidate predictors in terms of importance in determining the outcome. In this paper, our goal is to estimate the *marginal importance* of each predictor individually. The approach we present here is especially suited for situations in which ordinary least squares regression does not provide suitable models, i.e. we have a large number of candidate predictors and possible interactions between them.

From a practical point of view, quantifying the marginal importance of candidate predictor variables is important for the interpretation of the fitted models in view of the experimental follow-up. In order obtain this quantification, we define the importance as a parameter of interest and consider estimators of this parameter. This means that we focus on estimating the importance of a single variable in a model for a response variable Y and many candidate predictors, and not on model selection. We assume that an appropriate model selection procedure that produces a parsimonious model is given. It is important to bear in mind that this procedure is trading off bias and variance to fit a good model for Y based on (a subset of) the candidate predictors $X_1, \ldots, X_p$. When the true interest is in the marginal importance of a *single* variable, inference regarding this parameter based directly on the inferred model may be more biased than necessary (Van der Laan and Rubin, 2006). To overcome this, we use the framework of statistical inference for variable importance developed by Van der Laan (2006) and show how it can be applied to define and estimate variable importance of TFs in gene regulatory networks. This framework has previously been successfully applied to discover mutations that are clinically relevant to the treatment of HIV infection (Bembom, Petersen, Rhee, Fessel, Sinisi, Shafer, and Van der Laan, 2009).

The remainder of this paper is organized as follows. In Section 2 we give a definition of a variable importance measure (VIM) that makes sense within the context of gene regulatory networks and introduce three different estimators that we compare throughout this article. In Section 3 we study the behavior of these estimators in a simulation study. We show that the VIM we define represents a parameter that has an interesting biological interpretation by analyzing yeast gene expression in Section 4. Finally, we apply the VIM methodology to study the involvement of transcriptional regulators in determining gene expression of axonal growth-associated genes in Section 5. We conclude with a discussion in Section 6.

2 Methods

2.1 Marginal variable importance as a real-valued parameter

Suppose we observe a set of p predictors $X_1,\ldots,X_p$ and a response variable Y, all vectors of length n. We are interested in the marginal variable importance of X_j in determining Y, in a model where also possibly confounding predictors $X^*_{-j} = \{X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_p\}$ may be related to Y. Hence, when we model the effect of variable j, for $j = 1\ldots,p$ we consider the other variables X^*_{-j} as nuisance variables. For notational convenience we fix j and let $Z = X_j$ and $X^*_{-j} = X^*$. Within the VIM framework proposed by Van der Laan (2006), variable importance is modeled using a semi-parametric model that describes the effect of Z,X^* on Y as

$$\mathbb{E}(Y|Z,X^*) = m(Z,X^*|\beta) + g(X^*), \tag{1}$$

where $g(X^*)$ is an unspecified function of X^* and m is an *a priori* given model, which models the effect

$$m(Z = z,X^*|\beta) = \mathbb{E}[Y|Z = z,X^*] - \mathbb{E}[Y|Z = 0,X^*], \tag{2}$$

for all z. Based upon this specification, the following general definition of *marginal variable importance* ψ is suggested.

Definition Let models $\mathbb{E}(Y|Z,X^*)$ and $m(Z,X^*|\beta)$ as specified in (1) and (2) be given. The *marginal variable importance* (VIM) of variable Z at $Z = z$, denoted by $\psi(z)$, is defined as

$$\psi(z) = \mathbb{E}_{X^*}[m(z,X^*|\beta)]. \tag{3}$$

Here, we assume a linear model $m(Z,X^*|\beta) = \beta_j Z$ to model linear marginal effects. Furthermore, we consider $\psi = \psi(1)$ as the parameter of interest. The interpretation of this parameter is the expected change in Y for a unit change in Z while holding all other predictors fixed at their original values.

Example Let us consider an example. Suppose we have the following multiple linear regression model relating a response variable Y to a set Z, X^* of predictors

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_j Z + \ldots + \beta_p X_p + \varepsilon.$$

Within the framework introduced above, we write this as $\mathbb{E}(Y|Z,X^*) = m(Z,X^*|\beta) + g(X^*)$, where $m(Z,X^*|\beta) = \beta_j Z$ and

$$g(X^*) = \beta_0 + \beta_1 X_1 + \ldots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \ldots + \beta_p X_p.$$

In this case, the variable importance parameter is given by

$$\psi(z) = \mathbb{E}[Y|Z = z, X^*] - \mathbb{E}[Y|Z = 0, X^*] = \beta_j z,$$

and we focus on inference of $\psi(1) = \beta_j$.
**

For this example the function $g(X^*)$ consists exclusively of additive main effects of the variables X^*. In general, more complex functions $g(X^*)$ can be considered.

Other definitions of variable importance exist, but at present, a widely accepted standard methodology is absent. See e.g. Grómping (2007) for a discussion of theoretical and empirical properties of two key competing relative importance estimators for decomposition of the model variance R^2 of a the linear regression model. Chevan and Sutherland (1991) proposed "Hierarchical Partitioning" for more general univariate regression situations, and for non-parametric random forests regression (Breiman, 2001), the permutation importance is a practically useful measure of the impact of predictors (Strobl, Boulesteix, Kneib, Augustin, and Zeileis, 2008). Our practical interest here is in estimating the importance of biological predictors associated to gene expression and the general VIM definition in Equation (3) enables us to do this while using any appropriate model selection algorithm for estimating the full model $\mathbb{E}(Y|Z,X^*)$.

2.2 Targeted Maximum Likelihood

Targeted Maximum Likelihood estimation (TMLE) is a general framework that can be applied for the estimation of variable importance parameters. The theory behind TMLE was published by Van der Laan and Rubin (2006). Here, we provide a brief summary of the main idea. Classical maximum likelihood methods for estimating VIMs focus on estimation of the model $\mathbb{E}(Y|Z,X^*)$ by minimizing a global measure, such as the L_2-loss. When the primary interest is the estimation of one particular parameter of the data distribution, hence considering the remaining parameters as

nuisance parameters, an estimator that has smaller bias and variance for the parameter of interest would be preferred. TMLE relies on the following factorization of the likelihood of the observed data $O = (Y, Z, X^*)$

$$\mathbb{L}(O) = P(Y|Z,X^*)P(Z|X^*)P(X^*).$$

Standard approaches to VIM estimation rely only upon estimation of $\mathbb{E}(Y|Z,X^*)$, resulting in an estimate that represents a good bias-variance trade-off for the full regression $\mathbb{E}(Y|Z,X^*)$. However, for the parameter of interest, this estimate may be biased unnecessarily. TMLE involves estimation of $P(Z|X^*)$ as well and updates an initial estimate of $\mathbb{E}(Y|Z,X^*)$ by maximizing the likelihood in a direction which corresponds to the best estimate of the parameter of interest ψ. It was shown by Van der Laan and Rubin (2006) that when either $\mathbb{E}(Y|Z,X^*)$ or $\mathbb{E}(Z|X^*)$ are specified correctly, the targeted maximum likelihood estimator is consistent and asymptotically normal. Furthermore, if both models are specified correctly, it is efficient. In practice, the overall quality of the estimator depends on good estimates of $\mathbb{E}(Y|Z,X^*)$ and $\mathbb{E}(Z|X^*)$.

2.3 Estimation of variable importance

We now introduce two closely related estimators of the VIM parameter as defined in Equation (3). As is clear from the model specification in Equation (1), estimation of VIM requires estimation of $\mathbb{E}(Y|Z,X^*)$. The present context of gene regulatory networks allows us to use our model selection tool `GEMULA` (Geeven, Van Kesteren, Smit, and De Gunst, 2012) for this, which we first briefly describe here. For given Z, X^*, `GEMULA` performs a prioritization step that ranks the variables X^* based on their association with the response Y, thus producing a ranking $X^*_{r(1)} \ldots X^*_{r(p-1)}$. Then, `GEMULA` fits a model of the form

$$\mathbb{E}(Y|Z,X^*) = \beta_z Z + g(X^*), \tag{4}$$

where $g(X^*)$ are candidate terms that are allowed in the model. These terms are determined through the specification of a tuning parameter $\gamma = (\gamma_1, \gamma_2)$, where γ_1 represents the maximum allowed order of interactions between terms in the models and γ_2 the maximum number of candidate terms allowed in the model. For instance, when $\gamma = (1, 150)$, we have

$$g(X^*) = g_\gamma(X^*) = \beta_1 X^*_{r(1)} + \ldots + \beta_{150} X^*_{r(150)},$$

and when we set $\gamma = (3, 150)$, then

$$\begin{aligned} g_\gamma(X^*) &= \beta_1 X^*_{r(1)} + \ldots + \beta_9 X^*_{r(9)} \\ &\quad + \beta_{10} X^*_{r(1)} X^*_{r(2)} + \ldots + \beta_{45} X^*_{r(8)} X^*_{r(9)} \\ &\quad + \beta_{46} X^*_{r(1)} X^*_{r(2)} X^*_{r(3)} + \ldots + \beta_{129} X^*_{r(7)} X^*_{r(8)} X^*_{r(9)}, \end{aligned}$$

since including all 1st and 2nd order interactions between more than 9 predictors would exceed 150, which is the maximum number or terms allowed. For a given γ, `GEMULA` uses the lasso to fit an entire path of penalized coefficient estimates $\beta^\gamma_\lambda = (\beta^\gamma_{\lambda z}, \beta^\gamma_{\lambda 1}, \ldots, \beta^\gamma_{\lambda M(\gamma)})$ in model (4) for a range of shrinkage parameters $\lambda \in \Lambda$, where $M(\gamma)$ is the number of terms in $g_\gamma(X^*)$. We select the amount of shrinkage λ, and hence a corresponding estimate β^γ_λ, using the finite sample corrected version (Sugiura, 1978) of the Akaike information criterion (AIC). This produces a L_1-*penalized* VIM (lVIM) estimate of the parameter β_z in Equation (4). We denote this lVIM estimate by $\hat{\psi}_l$. Instead of using the lasso, we can also use the elastic net (Zou and Hastie (2005)) for model fitting. The elastic net is a regularization and model selection method that combines L_1 and L_2 regularization through a penalty $J(\beta) = \alpha \|\beta\|^2 + (1-\alpha)\|\beta\|_1$, where $0 < \alpha \leq 1$. The additional L_2-penalty encourages grouping of highly correlated predictors and stabilizes the L_1-regularization path (Zou and Hastie, 2005, Wang, Zhu, and Zou, 2006). For comparison we include an estimate $\hat{\psi}_e$ of ψ based on the elastic net with parameter $\alpha = 0.2$ and denote it as eVIM.

The Targeting Step (tVIM)

Let an initial fit of a model M_0 for $\mathbb{E}(Y|Z,X^*)$ and a fit of a model M_G for $\mathbb{E}(Z|X^*)$ be given. Below, we give the steps required for the computation of tVIM. For more details, we refer to Van der Laan (2006), Van der Laan and Rubin (2006).

1. Calculate a covariate $r(Z,X^*) = Z - \hat{Z}^{M_G}$, where $\hat{Z}^{M_G}$ are the fitted responses obtained from the model M_G for $\mathbb{E}(Z|X^*)$.
2. Compute the vector of fitted response values $\hat{Y}^{M_0}$ according to the fitted model M_0 for $\mathbb{E}(Y|Z,X^*)$.
3. Regress Y on $r(Z,X^*)$ using $\hat{Y}^{M_0}$ as an *offset* and denote the estimated regression coefficient by $\hat{\varepsilon}$. An offset is a term that can be added to a linear model and that is treated as an *a priori* known term, for which no coefficient needs

to be estimated. The offset is subtracted from the response prior to fitting. The estimate $\hat{\varepsilon}$ can be obtained by standard OLS regression, using a model without an intercept term but *with* the mentioned offset.

4. Update the initial lVIM estimate $\hat{\psi}_l$ to obtain the tVIM estimate $\hat{\psi}_t$ as

$$\hat{\psi}_t = \hat{\psi}_l + \hat{\varepsilon}.$$

3 Simulation study

In this section, we conduct a simulation study in order to compare the estimation of VIMs using the lVIM, eVIM and tVIM estimators described in Section 2.1. Because the VIM defined in equation (3) in Section 2.1 defines a linear effect, we use a linear model (that we derive from real experimental data, see Appendix for details) and make the comparison on data generated by this model. The main purpose is to show the effect of the targeting step on the performance of the tVIM estimator and to get some insight into its behavior. Note that the pilot model described in the Appendix is designed primarily to study models relating binding affinities of DNA binding TFs to observed variation in gene expression in the context we consider in this paper. Hence, simulations using this model provides us with perspective on the potential of VIM estimation for the identification and ranking (based on importance) of predictors associated to variation in gene expression. Here, we consider a data generating model that contains the linear main effects for all 33 predictors in the pilot model. We use a set of 123 `TRAP MRM` predictors (see Geeven et al. (2012) and the Appendix) as candidate predictors. Hence, only the 33 predictors present in the pilot model correspond to "truly important" predictors, i.e. predictors with a non-vanishing regression coefficient. For $j = 1,\ldots,33$, we let β_j represent the parameter of interest, i.e. the true VIM parameter ψ_j corresponding to predictor j. The response variable Y is generated as

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{33} X_{33} + \varepsilon, \tag{5}$$

where $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_n) \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$. In the simulations we consider here, we use $n = 790$, because the pilot model was derived by analyzing 790 cell cycle genes and we consider this to be a fairly typical value for a set of regulated genes. We set $\sigma^2 = 0.26$, which corresponds to a setting with high noise variance. On data simulated according to model (5), we compare the performance of the lVIM, eVIM and tVIM estimators, based on 1000 independent simulation runs. We also compare the lVIM, eVIM and tVIM estimates to estimates of the regression coefficients obtained with OLS. Hence, in each independent simulation run we record the estimated VIM according to the following four estimators.

1. **mOLS**. The first method we use to estimate the VIM for each candidate predictor j, for $j = 1, \ldots, 123$, is based on an OLS fit of the model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{123} X_{123} + \varepsilon. \quad (6)$$

   The estimate of β_j obtained from the resulting fit is recorded as the mOLS estimate of ψ_j.
2. **lVIM**. We use `GEMULA` with $\gamma = (1, 123)$ to obtain the lVIM estimate. With this setting, the model selected by `GEMULA` corresponds to an L_1-penalized lasso fit of model (6). The estimate of β_j that is obtained as the estimated regression coefficient corresponding to predictor j in the model selected by `GEMULA` is recorded as the lVIM estimate of ψ_j.
3. **eVIM**. We generate the candidate predictors exactly similar as for lVIM, but use the elastic net with parameter $\alpha = 0.2$ instead of the lasso to obtain a penalized estimate of ψ_j, which we denote by eVIM. For comparison, the amount of shrinkage is chosen by minimizing the analogous information theoretic criterion as for the lasso, i.e. AICc.
4. **tVIM**. The tVIM estimate is obtained by applying the steps described in Section 2.3 to the lVIM estimate of ψ_j in step 2. To estimate $\mathbb{E}(Z|X^*)$ we use `GEMULA` with $\gamma = (1, 15)$.

The predictors can be ranked according to their true importance $|\beta_j|$. Figure 1 contains box plots of the estimated VIMs according to the three methods for the 9 highest ranking predictors. The horizontal line in each plot represents the true VIM ψ_j. The plots in Figure 1 clearly illustrate how the targeting step works. It moves the shrunken, low variance (but biased) lVIM estimate in the direction of the true value of the parameter. As such, the resulting targeted VIM estimate represents a compromise between lVIM and mOLS. On average, it has a lower bias than lVIM but a higher variance. The quality of an estimator is a function of both bias and variance and a common way to quantify the distance between an estimator and a parameter of interest being estimated is to compute the mean square error (MSE), or its square root (RMSE). Table 1 contains RMSEs calculated for each of the three different estimators based on 200 simulations. From this Table, we conclude that for these "most important" predictors, tVIM gives the most accurate estimates in terms of RMSE. It is apparent that in most cases, the lVIM estimates are more biased than necessary. However, we note that this is not necessarily so for *all* predictors. As the effects ψ_j become smaller, at some point the negative impact of the additional variance in the estimates introduced by the targeting step overcomes the benefits of the reduction in bias. For smaller effects, shrinking them toward zero results in lower RMSEs. Hence, for predictors with very small effects, the eVIM estimates are most accurate. This is illustrated in Figure 2 and Table 2.

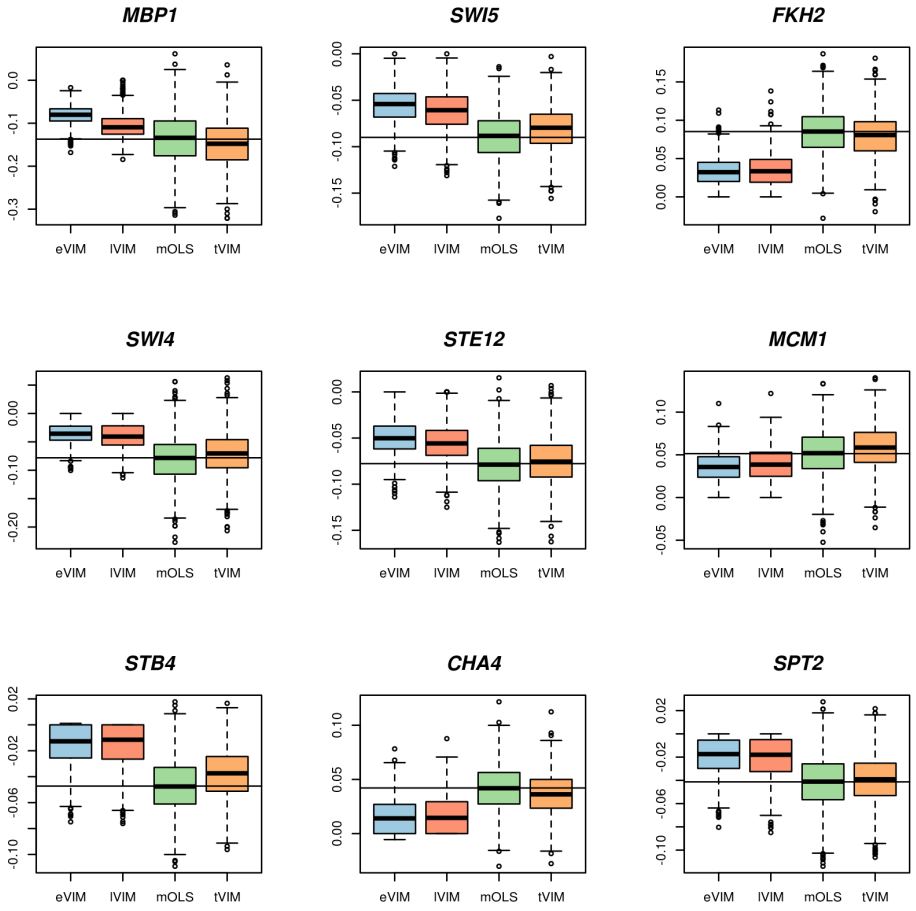

Figure 1: Boxplots of VIMs estimated using four different estimators for the 9 most important predictors in the simulation study.

From the results of the simulations we conclude that in general eVIM and lVIM perform very similarly, where perhaps eVIM represents a marginal improvement over lVIM for smaller effects. On the other hand tVIM differs from eVIM and lVIM, being less biased on average but producing estimates with higher variance. Summarizing, tVIM, eVIM and lVIM all provide good estimates of VIMs of interest on which rankings of marginal importance of predictors can be based, with none of the methods being superior in terms of RMSE across the entire range of effects. The main purpose of the simulations we present here is to illustrate the variable importance framework within a clearly interpretable context and to characterize the effect of the *targeting* step on the lVIM estimates. Although in this

Predictor	mOLS	eVIM	lVIM	tVIM	β_j
MBP1	0.0595	0.0602	**0.0440**	0.0561	-0.137
SWI5	**0.0251**	0.0394	0.0361	**0.0251**	-0.0900
FKH2	0.0300	0.0553	0.0545	**0.0285**	0.0852
SWI4	0.0422	0.0466	0.0452	**0.0408**	-0.0780
STE12	0.0270	0.0337	0.0309	**0.0257**	-0.0777
MCM1	0.0274	**0.0234**	0.0237	0.0264	0.0514
STB4	**0.0207**	0.0353	0.0357	0.0215	-0.0472
CHA4	0.0212	0.0297	0.0298	**0.0201**	0.0421
SPT2	0.0233	0.0273	0.0273	**0.0219**	-0.0413
FKH1	0.0306	0.0384	0.0383	**0.0278**	-0.0394

Table 1: RMSEs of four different VIM estimators for the 10 predictors with the highest effect size (β_j) in the simulation study. The smallest RMSE is indicated in boldface.

simulation example mOLS appears to yield reasonable VIM estimates too, it almost never outperforms eVIM, lVIM and tVIM. Moreover, within the general variable importance framework outlined in Section 2.1, it is not a natural estimator to consider. This framework enables us to estimate variable importance in the context of gene regulatory networks, where the general form of $\mathbb{E}(Y|Z,X^*)$ is unknown. We consider eVIM, lVIM and tVIM to be complementary and use them all to analyze real expression data in the following sections. The relative usefulness of the VIM estimates and rankings of predictors obtained using eVIM, lVIM and tVIM will become clear upon further validation and interpretation of the inferred results obtained on real gene expression data.

4 Validation on yeast gene expression data

In order to confirm that estimation of variable importance using lVIM, eVIM and tVIM estimators yields biologically relevant parameters when applied to real experimental data, we apply the outlined variable importance approach to yeast cell cycle gene expression data. We give a comprehensive analysis of the relative importance of different TFs that are associated to the observed variation in gene expression and focus on the dynamic activity of the TFs in time.

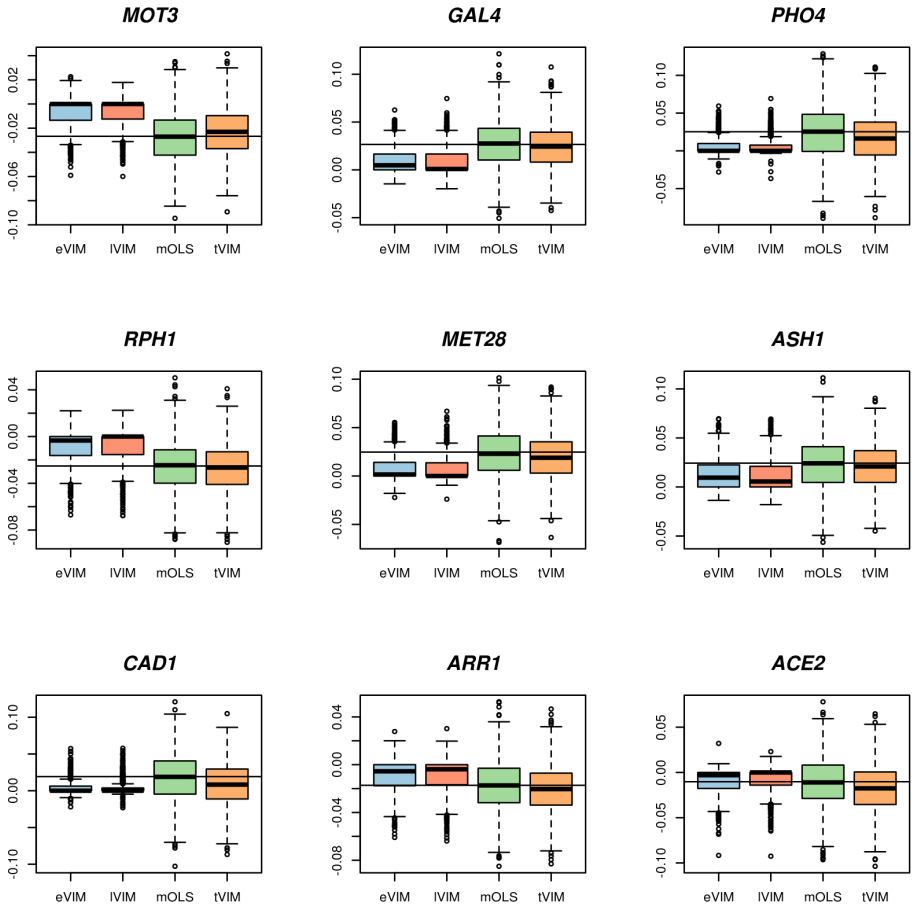

Figure 2: Boxplots of VIMs estimated using four different estimators for the 9 least important predictors in the simulation study.

Gene expression time-course profiles of synchronized yeast cultures progressing through the different stages of the cell cycle were measured by Spellman, Sherlock, Zhang, Iyer, Anders, Eisen, Brown, Botstein, and Futcher (1998). In their analysis of the 800 periodically expressed genes they identified as cell cycle regulated, Spellman *et al.* partitioned this set into five subsets based on the moment of peak expression during the cycle. In the following we use data from the entire set of experiments were α-factor arrest was used to synchronize the yeast cells. Expression was measured at 7 minute intervals up to 119 minutes after synchronization. Hence, the dataset we analyze consists of time-course gene expression profiles for all known yeast genes at 18 different time-points spanning two complete

Predictor	mOLS	eVIM	lVIM	tVIM	β_j
MOT3	0.0217	0.0223	0.0227	**0.0203**	-0.0267
GAL4	0.0252	**0.0208**	0.0215	0.0227	0.0266
PHO4	0.0364	**0.0219**	0.0225	0.0332	0.0253
RPH1	0.0220	**0.0202**	0.0208	0.0207	-0.0252
MET28	0.0262	**0.0200**	0.0205	0.0244	0.0247
ASH1	0.0257	**0.0182**	0.0195	0.0231	0.0243
CAD1	0.0325	**0.0173**	0.0178	0.0314	0.0193
ARR1	0.0212	**0.0142**	0.0148	0.0196	-0.0173
ACE2	0.0275	**0.0135**	0.0136	0.0265	-0.0103

Table 2: RMSEs of four different VIM estimators for the 10 predictors with the lowest effect size (β_j) in the simulation study. The smallest RMSE is indicated in boldface.

cell cycles. Figure 3 shows the average expression profiles of the 800 periodically expressed genes clustered by time of peak expression. In this plot the distinct cell cycle phases are indicated in boldface font. This plot clearly shows the periodicity of the gene expression response and the different moments of peak expression of the different clusters of genes. Transcriptional regulation of cell cycle periodic genes has been studied intensively and analysis of different sources of experimental data has identified various TFs that underlie the periodic patterns of gene expression Tsai, Lu, and Li (2005), Cokus, Rose, Haynor, Gronbech-Jensen, and Pellegrini (2006), Wu and Li (2008). Cokus et al. (2006) describe interactions between the primary or *canonical* cell cycle regulators `SWI4`, `SWI6`, `MBP1`, `FKH2`, `NDD1`, `MCM1`, `SWI5` and `ACE2` which are known to form complexes and regulate phase transitions in the cycle in a serial fashion. Tsai et al. (2005) identify a set of thirty putative cell cycle TFs. For nineteen of these there is strong evidence in the literature. The list of cell cycle TFs reported in Tsai et al. (2005) includes the eight canonical TFs discussed in Cokus et al. (2006). In the canonical model of transcriptional regulation of the cell cycle, the different primary regulators activate their targets at the different phases (`M/G1`, `G1/S` and `G2/M`) in the cell cycle. We investigate whether we can reconstruct the activities of these canonical TFs by estimating their marginal variable importance for the different cell cycle phases.

In order to identify the TFs that control the expression of these genes, we rank predictors based on their estimated marginal variable importance using the

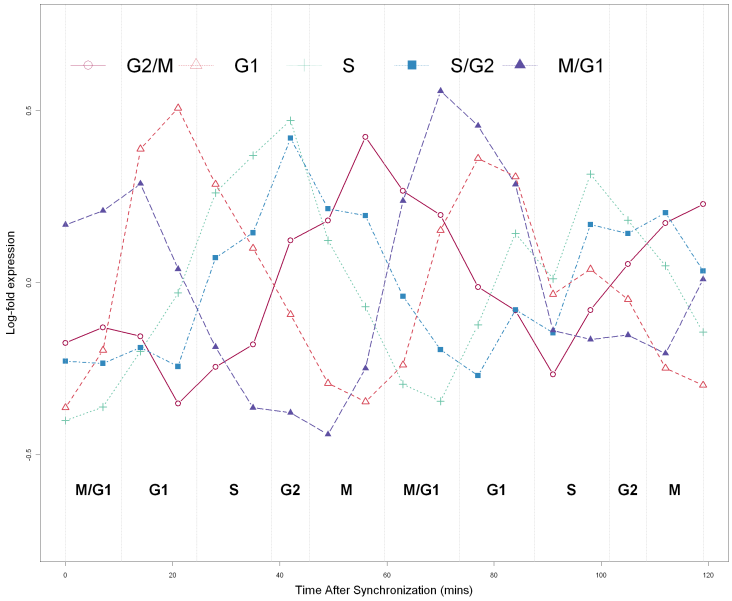

Figure 3: Observed gene expression across two complete cell cycles of 800 cell cycle regulated yeast genes, clustered by time of peak expression.

eVIM, lVIM and tVIM estimators. We again use the `TRAP MRM` predictors that are constructed using PFMs from 123 different yeast TFs derived from experimental binding data published by Macisaac, Wang, Gordon, Gifford, Stormo, and Fraenkel (2006). We estimate $\mathbb{E}(Y|Z,X^*)$ using `GEMULA` with parameter $\gamma = (2,250)$. The resulting fitted model is used to produce the lVIM estimate of ψ_j. To compute the tVIM estimate of ψ_j, we estimate $G(X^*) = \mathbb{E}(Z|X^*)$ using `GEMULA` with $\gamma = (1,15)$ and update the lVIM according to the steps described in Section 2.3. Figure 3 shows a large cluster of genes that peak 21 minutes following alpha synchronization, a time-point that lies within the `G1` phase of the cell cycle. Table 3 lists the highest ranked predictors and the estimated effect sizes according to eVIM, lVIM and tVIM for this time-point. A predictor is included in Table 3 if and only if it ranks among the 10 highest according to *either* eVIM, lVIM or tVIM. The order in which the predictors appear from top to bottom in Table 3 is determined by their tVIM rank. The top ranked predictors `MBP1` and `STB1` are both known transcriptional activators of cell cycle genes during the `G1` phase of the cycle in Tsai et al. (2005). The positive values for the estimate of the VIMs of `MBP1` and `STB1` at this time-point indeed agree with their known role as activators of genes during `G1`. Table 3 also identifies the canonical regulators `FKH2` and `ACE2`. Furthermore, the factor `SFP1` is a known regulator of `G2/M` cell cycle transitions (note the negative sign of the

Predictor	tVIM	tVIM rank	lVIM	lVIM rank	eVIM	eVIM rank
MBP1	0.203	1	0.199	1	0.175	1
STB1	0.092	2	0.081	2	0.065	2
SFP1	-0.082	3	-0.023	11	-0.021	14
FKH2	-0.073	4	-0.061	3	-0.058	3
HAC1	0.068	5	0.033	5	0.034	5
REB1	-0.056	6	-0.025	9	-0.023	11
SKO1	0.055	7	0.031	6	0.03	7
ACE2	-0.051	8	-0.026	8	-0.026	9
ASH1	-0.046	9	-0.028	7	-0.025	10
AZF1	-0.044	10	-0.034	4	-0.030	6
YAP3	-0.043	14	-0.025	10	-0.022	13
SWI4	0.035	19	0.011	23	0.03	8
SWI6	0.002	50	0	NA	0.056	4

Table 3: Top ranked predictors by tVIM, lVIM and eVIM. The response variable is observed gene expression of yeast cell cycle regulated genes 21 minutes after synchronization. The canonical cell cycle regulators are indicated in boldface and TFs belonging to the set of 19 known cell cycle TFs in Tsai et al. (2005) in italics.

estimated variable importance during G1) Cherry, Adler, Ball, Chervitz, Dwight, Hester, Jia, Juvik, Roe, Schroeder, Weng, and Botstein (1998) and also ASH1 and DIG1 are implicated in regulation of cell cycle genes according to Tsai et al. (2005).

Another important gene expression pattern is due to genes that peak at the transition from G2 to M phase, corresponding roughly to the time-point 56 minutes after synchronization (see Figure 3). The top ranked predictors for this time point are listed in Table 4. Selection and ranking of predictors was done as for Table 3. We find high positive marginal importances of the canonical factors FKH2 and MCM1, both linked to the activation of M and G2/M cell cycle genes respectively according to Tsai et al. (2005). Apart from MBP1 and MCM1, the top ranked predictors in Table 4 also include the canonical regulators SWI4, SWI5, ACE2 and FKH2. Note that FKH1, a TF that is part of the set of nineteen TFs with literature support for being important in cell cycle regulation according to Tsai et al. (2005) can only

Predictor	tVIM	tVIM rank	lVIM	lVIM rank	eVIM	eVIM rank
MBP1	-0.125	1	-0.11	1	-0.061	1
FKH2	0.089	2	0.04	6	0.034	7
SWI4	-0.079	3	-0.062	2	-0.04	5
MCM1	0.076	4	0.056	4	0.05	4
STE12	-0.071	5	-0.061	3	-0.054	3
SWI5	-0.068	6	-0.048	5	-0.056	2
PHO4	0.045	7	0.002	18	0.003	24
ACE2	-0.044	8	0	NA	-0.027	8
FKH1	-0.038	9	0	NA	0	NA
RDS1	0.035	10	0.018	8	0.015	10
PHD1	0.035	11	0.015	9	0.012	13
SWI6	-0.03	17	0	NA	-0.039	6
PDR3	0.024	22	0.018	7	0.015	11
GCN4	0.001	50	0.014	10	0.018	9

Table 4: Top ranked predictors by tVIM, lVIM and eVIM. The response variable is observed gene expression of yeast cell cycle regulated genes at 56 mins after synchronization. The canonical cell cycle regulators are indicated in boldface and TFs belonging to the set of 19 known cell cycle TFs in Tsai et al. (2005) in italics.

be identified using tVIM, and ACE2 only by eVIM and tVIM. Also note that the crucial M phase regulator FKH2 ranks second in the tVIM list and only sixth and seventh respectively in the lists produced using lVIM and eVIM. In contrast, we found no evidence in the literature for any specific cell cycle regulatory role for the TFs PDR3 and GCN4, which only receive high ranks according to lVIM and eVIM. Together, these findings illustrate the additional benefit of the targeting step and the tVIM estimator. The usefulness of the variable importance parameter as defined by (1) and (3) is further demonstrated in Figure 4. This plot shows the estimated marginal variable importance of the canonical cell cycle TFs MBP1, MCM1, FKH2 and SWI5 as a function of time in the succesive stages of the cell cycle. Most prominent is the clearly periodically varying importance of MBP1, peaking in the G1 phase. This is in good agreement with MBP1's known role as activator of cell cycle genes at the transition from G1 to S phase. Furthermore, the plots in Figure 4 suggest MCM1

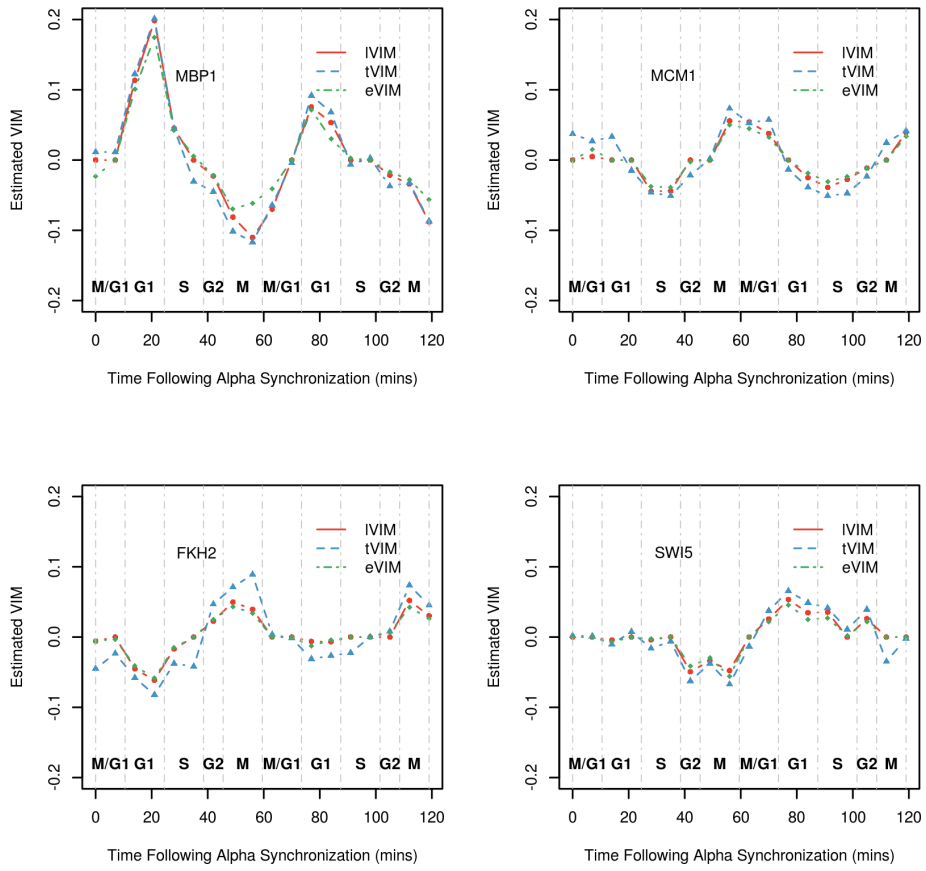

Figure 4: Plot of estimated VIMs of four canonical yeast cell cycle regulators across differents phases of the cell cycle.

and FKH2 as G2/M regulators and an involvement of SWI5 in the M/G1 transition. All of these findings are in agreement with what is known in the literature about the transcriptional effects of these TFs.

5 Estimation of VIM: an application

Here we apply the VIM methodology to estimate the variable importance of TFs in the gene regulatory network underlying neuronal outgrowth. As a cellular model we consider F11 cells (Platika, Boulos, Baizer, and Fishman, 1985). Upon stimulation with Forskolin, F11 cells acquire a neuronal phenotype, which results in the

outgrowth of neurites (Ghil, Kim, Lee, and Suh-Kim, 2000). We reanalyzed previously published genome-wide gene expression time course profiles of F11 cells measured at four time-points following Forskolin stimulation (MacGillavry, Cornelis, van der Kallen, Sassen, Verhaagen, Smit, and Kesteren, 2011). Previously, we showed that the entire set of Forskolin responsive genes, i.e. genes differentially expressed in response to stimulation, can be further divided into groups of "early" and "late" responsive genes (Geeven et al., 2012). We applied `GEMULA` to infer two models for the early responsive genes at the 2 hour and 4 hour time-point and two models for the late responsive genes at the 24 hour and 48 hour time-point. We again distinguish between these two groups. For the early responsive genes, we use `GEMULA` with $\gamma = (2, 500)$ for the estimation of $\mathbb{E}(Y|Z, X^*)$ and for the late responsive genes, which is a bigger set, we use `GEMULA` with $\gamma = (2, 700)$. For the estimation of $\mathbb{E}(Z|X^*)$ we use $\gamma = (2, 1, 110)$.

Predictor	tVIM	rank	lVIM	rank	eVIM	rank
V.CEBPDELTA.Q6	0.063	1	0.047	1	0.04	1
V.OCT1.03	0.056	2	0.031	3	0.026	3
V.PAX4.02	0.053	3	0.017	11	0.016	9
V.CIZ.01	0.051	4	0.02	8	0	NA
V.YY1.Q6.02	-0.051	5	-0.012	14	-0.011	12
V.CP2.02	0.047	6	0.026	5	0.02	5
V.CREB.Q4.01	0.043	7	0.039	2	0.034	2
V.AP1.Q4.01	0.041	8	0.02	9	0.01	13
V.DR1.Q3	-0.039	9	0	NA	0	NA
V.LEF1.Q2.01	0.037	10	0.006	23	0.007	18
V.PBX.Q3	-0.035	11	-0.02	7	-0.018	7
V.AREB6.02	-0.034	13	-0.027	4	-0.023	4
V.VJUN.01	0.026	20	0.017	10	0.017	8
V.E2F.Q6.01	0.009	49	0.021	6	0.02	6

Table 5: Top ranked predictors by lVIM, tVIM and eVIM. Response variable Y represents log-fold gene expression in cultured F11 cells of early responsive genes at 2h after Forskolin stimulation with respect to control.

Predictor	tVIM	rank	lVIM	rank	eVIM	rank
V.E2F.Q6.01	-0.123	1	-0.1	1	-0.065	1
V.MYB.Q3	-0.081	2	-0.013	16	-0.002	28
V.LRF.Q2	0.077	3	0	NA	0.006	22
V.AP1.Q4.01	0.066	4	0.016	9	0.006	21
V.COUP.DR1.Q6	0.057	5	0.015	10	0.01	15
V.GEN.INI3.B	0.056	6	0.008	20	0.011	13
V.E2A.Q2	0.056	7	0.018	6	0.014	10
V.EBF.Q6	0.054	8	0.033	3	0.027	3
V.OCT1.Q5.01	-0.05	9	-0.001	29	-0.003	26
V.NKX3A.01	-0.045	10	-0.007	23	-0.008	19
V.POU6F1.01	-0.045	11	-0.025	4	-0.017	6
V.PPAR.DR1.Q2	0.043	12	0.017	7	0.015	8
V.PAX4.03	0.043	13	0.045	2	0.028	2
V.PPARA.01	0.042	14	0.015	11	0.014	9
V.P300.01	0.039	18	0.017	8	0.017	7
V.VDR.Q3	0.039	19	0.025	5	0.022	5
V.E2F.03	-0.002	54	-0.001	30	-0.025	4

Table 6: Top ranked predictors by lVIM, tVIM and eVIM. Response variable Y represents log-fold gene expression in cultured F11 cells of late responsive genes at 24h after Forskolin stimulation with respect to control.

The results for the first time-point at two hours following Forskolin stimulation are presented in Table 5. A total of 212 different predictors were considered. A predictor is included in Table 5 if and only if it ranks among the 10 highest according to *either* eVIM, lVIM or tVIM. The order in which the predictors appear from top to bottom in Table 5 is determined by their tVIM rank. Note the high ranking of the binding site motifs `V.CREB.Q4.01` and `V.AP1.Q4.01`. Activation of CREB is known to be induced by Forskolin stimulation of F11 cells (MacGillavry, Stam, Sassen, Kegel, Hendriks, Verhaagen, Smit, and Van Kesteren, 2009) and the role of TFS binding to the motifs `V.CREB.Q4.01`, `V.AP1.Q4.01` and `V.VJUN.01` binding site motifs in driving gene expression in biological models

in neuronal regeneration is well established (Gao, Hou, Bryson, Barco, Nikulina, Spencer, Mellado, Kandel, and Filbin, 2004, Seijffers, Mills, and Woolf, 2007). We report the results for two later time points in Table 6 and 7. According to these tables, there is a strong repression of genes by the known cell cycle regulator E2F. Since cell cycle arrest and neurogenesis are highly coordinated and interactive processes (Ohnuma *et al.* Ohnuma and Harris (2003)), the involvement of E2F in regulation of genes in Forskolin stimulated F11 cells is plausible. Among the top 10 ranked TFBS motifs at the 24 hours and 48 hours time-point are V.PPARA.01 and V.PPAR.DR1.Q2. The consensus sequence of the TFBSs corresponding to this motif is recognized by TFs from the family of *peroxisome proliferator-activated receptors* (PPARs). In earlier work, we predicted PPARs to regulate genes involved in neuronal differentiation based on analysis of *in vivo* gene expression data from rat DRG neurons in response to injury using LLM3D (Geeven, MacGillavry, Eggers, Sassen, Verhaagen, Smit, De Gunst, and Van Kesteren, 2011), a method that uses log-linear modeling to detect enrichment of TF binding sites in functionally homogeneous sets of genes. There, we also described the validation of the effect of PPARγ on regulation of genes involved in neuronal differentiation. Our findings here provide further support for our claim that PPARγ is an important transcriptional regulator in neuronal regeneration. In addition to V.PPARA.01 in Table 6 and Table 7 we find V.EBF.Q6. This motif is bound by *early B-cell factor* (EBF) TFs. Garel *et al.* (Garel, Marín, Mattéi, Vesque, Vincent, and Charnay (1997)) find that EBFs are potentially involved in neuronal differentiation in the developing CNS. In Garcia-Dominguez, Poquet, Garel, and Charnay (2003), Dominguez *et al.* find that EBFs appear to be master controllers of neuronal differentiation and migration, coupling them to cell cycle exit and earlier steps of neurogenesis. A review by Liberg *et al.* (Liberg, Sigvardsson, and Akerblad (2002)) discusses the role of EBFs as regulators of differentiation in embryonic neural development. This review also describes interactions between *CCAAT/ enhancer-binding proteins* (C/EBPs), *sterol regulatory binding protein 1* (SREBP1), PPARγ and EBFs in adipocyte development. Interestingly, we also identify a C/EBP motif, V.CEBPDELTA.Q6 at the 2 hour and 4 hour time-point (not shown). It was shown in MacGillavry et al. (2011) that both C/EBPα and C/EBPβ are transcriptional targets of CREB and knockdown of C/EBPα and C/EBPβ significantly reduced neurite outgrowth *in vitro*. Another interesting result is the high ranking of the TRANSFAC motif V.TST1.01 in Table 7. This motif is bound by the *suppressed cAMP-inducible POU protein* (Scip alias Tst-1). Gondré *et al.* Gondre, Burrola, and Weinstein (1998) have studied the function of Scip in schwann cells, which are glia (non-neuronal cells) in the peripheral nervous system. The expression of Scip is required for the establishment of normal nerves and it is re-expressed during regeneration. Furthermore, regeneration and hypertrophy of axons and myelin is markedly accelerated in transgenic

Predictor	tVIM	rank	lVIM	rank	eVIM	rank
V.E2F.Q6.01	-0.158	1	-0.129	1	-0.092	1
V.TST1.01	-0.129	2	-0.010	22	-0.010	24
V.MYB.Q3	-0.120	3	-0.038	3	-0.046	2
V.LRF.Q2	0.096	4	0	NA	0.003	44
V.AP1.Q4.01	0.067	5	0.020	9	0.017	15
V.E2A.Q2	0.063	6	0.015	12	0.015	17
V.GEN.INI3.B	0.061	7	0.015	17	0.019	10
V.PAX4.03	0.056	8	0.053	2	0.036	3
V.OCT1.Q5.01	-0.055	9	-0.009	26	-0.010	28
V.EBF.Q6	0.055	10	0.036	4	0.032	4
V.SP3.Q3	0.053	11	0.031	5	0.024	6
V.PPARA.01	0.050	12	0.029	6	0.020	9
V.POU6F1.01	-0.050	13	-0.027	8	-0.022	7
V.MRF2.01	-0.046	15	-0.020	10	-0.021	8
V.CETS1P54.02	-0.041	18	-0.016	11	-0.015	18
V.VDR.Q3	0.040	19	0.027	7	0.030	5

Table 7: Top ranked predictors by lVIM, tVIM and eVIM. Response variable Y represents log-fold gene expression in cultured F11 cells at 48 hours after Forskolin stimulation with respect to control.

mice expression a Δ`Scip` transgene Gondre et al. (1998). Although the fact that we identify `Tst-1` as an important regulator of neuronal F11 cells may be surprising, it may be interesting to further study the role of this TF in neurons. Interactions between neurons and glial cells play important roles in regulating key events of development and regeneration of the CNS. Also, Table 6 and 7 list another `POU`-domain motif, `V.POU6F1.01`. The various members of the `POU` family have a wide variety of functions, all of which are related to the development of an organism.

Altogether the results we present here identify several known and some putative novel DNA binding motifs that correspond to TFs which are likely to be important in the transcriptional regulatory network underlying neuronal regeneration. The VIM parameters allow us to estimate the variable importance for several

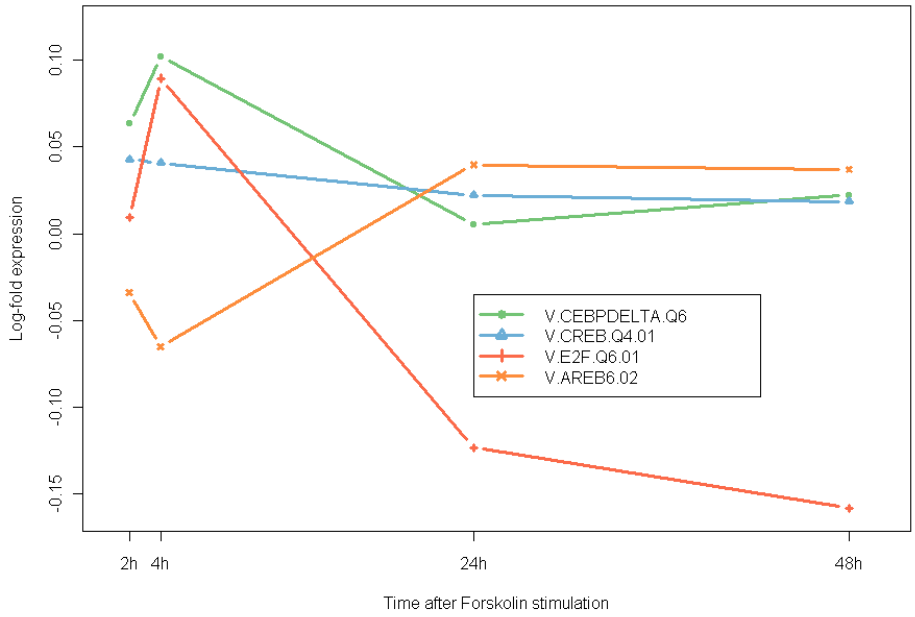

Figure 5: Plot of tVIM estimates versus time following Forskolin stimulation for several `TRAP TF11` predictors associated to gene expression changes in F11 cells in response to Forskolin stimulation.

highly ranked `TRAP TF11` predictors at several time-points to get some insight into the dynamic activity of the corresponding TFs, as we did in the analysis of yeast cell cycle gene expression data in Section 4. The plot is drawn in Figure 5.

6 Discussion

The application of the variable importance estimation framework that we consider in this paper to analyze real experimental gene expression data, provides biologically meaningful results. In particular, when genomewide time-course profiles of gene expression are available, it allows us to identify parameters that can be interpreted as representations of dynamic activity of transcriptional regulators underlying the observed patterns of gene expression. In Section 4 we validated the results we obtained on yeast cell cycle data against the literature on transcriptional regulation of the yeast cell cycle. This literature is largely based on analysis of *in vivo* binding data, such as ChIP-chip assays. Of course, when such binding data is

available, it can be used to replace or complement the surrogate predictors we use, i.e. TRAP predictors that represent binding affinities derived *in silico* according to a biophysical model of DNA binding by TFs. However, our main goal here is to illustrate how to infer as much as possible about context specific regulatory effects of TFs on observed gene expression in absence of such data. The results of the analysis of yeast gene expression data in Section 4 show that the use of surrogate binding affinities as obtained using TRAP enables us to reconstruct the time dependent effect of several known cell cycle TFs such as MBP1, MCM1, FKH2 and SWI5 remarkably well.

In order to be able to use experimental *in vivo* binding data, these data should be obtained under the same experimental conditions under which the gene expression was measured. For mammalian gene expression experiments, such experimental binding data is typically only available for a couple of TFs or even completely lacking. This is also the case for the gene expression data from the *in vitro* biological model of neuronal regeneration we analyzed in Section 5. We identified known and putative novel TFs that are associated to patterns of early and late gene expression changes in F11 cells in response to Forskolin stimulation. The sign of the VIM parameter can be used to distinguish between transcriptional activators and repressors of gene expression. For instance, we estimated a positive value for the VIM of V.E2F.Q6.01 at 2h following Forskolin stimulation and negative values, indicating repression of genes, for V.E2F.Q6.01 at the two later time-points. Such information on dynamic activity of TFs (see Figure 5) is important for understanding the evolution of transcriptional regulatory networks in time.

Appendix

In this Appendix we describe the model, hereafter called the *pilot model*, that we use for conducting simulations and generating data for the comparison of different VIM estimators that we are interested in. This model is derived from real experimental gene expression data and can be viewed as an approximation of the ”true” data generating model of a gene regulatory network. To this end, we fit a linear model with a stepwise variable selection algorithm to real yeast gene expression data and the resulting model will be our pilot model from which the artificial gene expression data will be simulated.

The pilot model

Yeast gene expression data from the cell cycle study performed by Spellman *et al.* was introduced in Section 4. Here, we consider the α-factor arrest experiments,

which contain measurements of all yeast genes at 18 different time points following synchronization, spanning three complete cell cycles (periods). In the original study by Spellman *et al.*, 800 yeast genes were identified as being cell cycle regulated. The expression profiles of these 800 genes display a clearly distinguishable periodic pattern, which is well known to be governed by a number of different transcription factors. We consider the expression of the 800 cell cycle regulated genes at 56 minutes following α-factor arrest. The microarrays used to measure cell cycle expression are two-channel arrays, hence the observed gene expression values correspond to log-ratios of expression at 56 minutes compared to control. The 56 minute timepoint in the Spellman *et al.* cell cycle data corresponds roughly to the cell cycle phase just after the transition from `G2` to `M`. The motivation for this particular time-point is that the transition from `G2` to `M` is known to be coordinated transcriptionally. The Spellman gene expression data contain some missing data, resulting from spots on the microarray for which no accurate log-fold expression ratios could be obtained, but given the high degree of correlation between periodically co-expressed transcripts, the missing values can be estimated in a reliable way. We use the `KNNImpute` algorithm developed by Troyanskaya *et al.* Troyanskaya, Cantor, Sherlock, Brown, Hastie, Tibshirani, Botstein, and Altman (2001) to impute missing gene expression values.

From the experimentally derived DNA binding sites published by (Macisaac et al., 2006), we extract 123 different position frequency matrices representing models of the DNA sequences bound by the different transcription factor proteins. We use the `TRAP` (Roider, Kanhere, Manke, and Vingron, 2007) tool to calculate DNA binding affinities for binding to the genomic DNA sequences from 1 bp to 1000 bp directly upstream of the cell cycle regulated genes. The genomic sequences were obtained from SGD Cherry et al. (1998). For 10 out of the 800 cell cycle genes, we could not match the IDs to IDs of the genomic sequences from SGD, which brings the total sample down to 790. The resulting predictors $X_1,\ldots,X_{123}$ thus represent binding affinities of 123 yeast DNA binding TFs.

In order to fit the pilot model, we first need to define a set of appropriate linear candidate models $\mathscr{M}$ to consider. It is known that cooperation between TFs is important for cell cycle gene regulation and others have reported pairwise interactions between yeast cell cycle TFs Zhang, Wildermuth, and Speed (2008), Das et al. (2004). We therefore focus on identifying main effects and effects corresponding to interaction effects between the predictors. With 123 variables, there are 7503 possible candidate pairwise effects to consider. Since we expect only a subset of predictors to be truly associated to Y, we do a univariate-screening to select the predictors most strongly associated to Y univariately. We allow only interaction terms between these predictors in candidate models. For all candidate predictors X_j, for $j = 1,\ldots,123$, we calculate t-statistics indicating the significance of the estimate

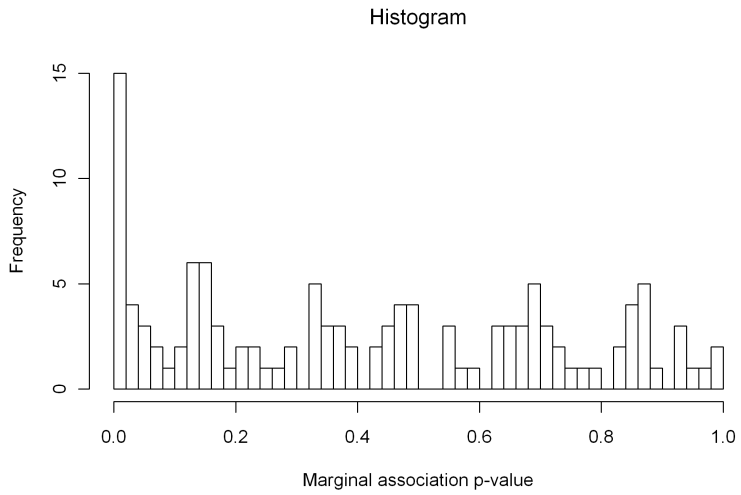

Figure 6: Marginal association p-values of all 123 different TRAP predictors.

of the regression coefficient β_j in the model $Y = \beta_0 + \beta_j X_j + \varepsilon$. The p-values corresponding to the tests $H_{0j} : \beta_j = 0$ can be used to rank the predictors. Figure 6 contains a histogram plot of all 123 marginal p-values. The lowest observed p-value (unadjusted for 123 tests) is 2.1×10^{-18} which indicates strong evidence of association. There are 21 predictors with a marginal unadjusted p-value smaller than 0.05. Based on these results, we fit our pilot model by limiting the set of candidate predictor terms to

1. main effects for all predictor variables X_j, $j = 1, \ldots, 123$,
2. pairwise interaction effects for the 25 most strongly univariately associated predictor variables.

This brings the total number of candidate terms to $123 + 300 = 423$. We use the `step()` function in `R` r (2005) to perform model selection. This function implements a greedy stepwise search strategy that considers both forward and backward moves, starting from an initial model that contains an intercept term only. We use the AIC to evaluate and compare visited models in the stepwise search. Our resulting pilot model contains 57 terms, 33 main effects and 23 first-order (pair-wise) interactions and an intercept term. The observed multiple (unadjusted) R^2 for the pilot model is 0.41. The estimate of the variance of the noise is $\hat{\sigma}^2 = 0.15$.

References

Bembom, O., M. L. Petersen, S.-Y. Rhee, W. J. Fessel, S. E. Sinisi, R. W. Shafer, and M. J. Van der Laan (2009): "Biomarker discovery using targeted maximum-likelihood estimation: Application to the treatment of antiretroviral-resistant hiv infection," *Statistics in Medicine*, 28, 152–172, URL http://dx.doi.org/10.1002/sim.3414.

Breiman, L. (2001): "Random forests," *Machine Learning*, 45, 5–32, URL http://dx.doi.org/10.1023/A:1010933404324.

Bussemaker, H., H. Li, and E. Siggia (2001): "Regulatory element detection using correlation with expression," *NATURE GENETICS*, 27, 167–171.

Cherry, J. M., C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng, and D. Botstein (1998): "Sgd: Saccharomyces genome database." *Nucleic acids research*, 26, 73–79, URL http://dx.doi.org/10.1093/nar/26.1.73.

Chevan, A. and M. Sutherland (1991): "Hierarchical partitioning," *The American Statistician*, 45, pp. 90–96, URL http://www.jstor.org/stable/2684366.

Cokus, S., S. Rose, D. Haynor, N. Gronbech-Jensen, and M. Pellegrini (2006): "Modelling the network of cell cycle transcription factors in the yeast saccharomyces cerevisiae," *BMC Bioinformatics*, 7, 381, URL http://www.biomedcentral.com/1471-2105/7/381.

Das, D., N. Banerjee, and M. Q. Zhang (2004): "Interacting models of cooperative gene regulation," *Proceedings of the National Academy of Sciences of the United States of America*, 101, 16234–16239, URL http://www.pnas.org/content/101/46/16234.abstract.

Gao, Y., J. Hou, J. Bryson, A. Barco, E. Nikulina, T. Spencer, W. Mellado, E. R. Kandel, and M. T. Filbin (2004): "Activated creb is sufficient to overcome inhibitors in myelin and promote spinal axon regeneration in vivo," *Neuron*, 44, 609–621.

Garcia-Dominguez, M., C. Poquet, S. Garel, and P. Charnay (2003): "Ebf gene function is required for coupling neuronal differentiation and cell cycle exit," *Development*, 130, 6013–6025, URL http://dev.biologists.org/content/130/24/6013.abstract.

Garel, S., F. Marín, M. Mattéi, C. Vesque, A. Vincent, and P. Charnay (1997): "Family of Ebf/Olf-1-related genes potentially involved in neuronal differentiation and regional specification in the central nervous system." *Developmental Dynamics*, 210, 191–205.

Geeven, G., H. D. MacGillavry, R. Eggers, M. M. Sassen, J. Verhaagen, A. B. Smit, M. C. M. De Gunst, and R. E. Van Kesteren (2011): "LLM3D: a log-linear modeling-based method to predict functional gene regulatory interactions from genome-wide expression data," *Nucleic Acids Research*, URL http://dx.doi.org/10.1093/nar/gkr139.

Geeven, G., R. E. Van Kesteren, A. B. Smit, and M. C. M. De Gunst (2012): "Identification of context-specific gene regulatory networks with gemulagene expression modeling using lasso," *Bioinformatics*, 28, 214–221, URL http://bioinformatics.oxfordjournals.org/content/28/2/214.abstract.

Ghil, S.-H., B.-J. Kim, Y.-D. Lee, and H. Suh-Kim (2000): "Neurite outgrowth induced by cyclic amp can be modulated by the a subunit of go," *Journal of Neurochemistry*, 74, 151–158, URL http://dx.doi.org/10.1046/j.1471-4159.2000.0740151.x.

Gondre, M., P. Burrola, and D. E. Weinstein (1998): "Accelerated Nerve Regeneration Mediated by Schwann Cells Expressing a Mutant Form of the POU Protein SCIP," *J. Cell Biol.*, 141, 493–501, URL http://jcb.rupress.org/cgi/content/abstract/141/2/493.

Grómping, U. (2007): "Estimators of relative importance in linear regression based on variance decomposition," *The American Statistician*, 61, 139–147, URL http://pubs.amstat.org/doi/abs/10.1198/000313007X188252.

Liberg, D., M. Sigvardsson, and P. Akerblad (2002): "The EBF/Olf/Collier Family of Transcription Factors: Regulators of Differentiation in Cells Originating from All Three Embryonal Germ Layers," *Mol. Cell. Biol.*, 22, 8389–8397, URL http://mcb.asm.org.

MacGillavry, H. D., J. Cornelis, L. R. van der Kallen, M. M. Sassen, J. Verhaagen, A. B. Smit, and R. E. V. Kesteren (2011): "Genome-wide gene expression and promoter binding analysis identifies nfil3 as a repressor of c/ebp target genes in neuronal outgrowth," *Molecular and Cellular Neuroscience*, 46, 460 – 468, URL http://www.sciencedirect.com/science/article/B6WNB-51JPWT8-4/2/df56199371a321743dd66962b7c53afc.

MacGillavry, H. D., F. J. Stam, M. M. Sassen, L. Kegel, W. T. J. Hendriks, J. Verhaagen, A. B. Smit, and R. E. Van Kesteren (2009): "NFIL3 and cAMP Response Element-Binding Protein Form a Transcriptional Feedforward Loop that Controls Neuronal Regeneration-Associated Gene Expression," *J. Neurosci.*, 29, 15542–15550, URL http://www.jneurosci.org/cgi/content/abstract/29/49/15542.

Macisaac, K., T. Wang, D. B. Gordon, D. Gifford, G. Stormo, and E. Fraenkel (2006): "An improved map of conserved regulatory sites for saccharomyces cerevisiae," *BMC Bioinformatics*, 7, 113+, URL http://dx.doi.org/10.1186/1471-2105-7-113.

Ohnuma, S.-i. and W. A. Harris (2003): "Neurogenesis and the cell cycle," *Neuron*, 40, 199–208.

Platika, D., M. H. Boulos, L. Baizer, and M. C. Fishman (1985): "Neuronal traits of clonal cell lines derived by fusion of dorsal root ganglia neurons with neuroblastoma cells," *Proceedings of the National Academy of Sciences*, 82, 3499–3503, URL http://www.pnas.org/content/82/10/3499.abstract.

r (2005): *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org.

Roider, H. G., A. Kanhere, T. Manke, and M. Vingron (2007): "Predicting transcription factor affinities to DNA from a biophysical model," *Bioinformatics*, 23, 134–141, URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/2/134.

Seijffers, R., C. D. Mills, and C. J. Woolf (2007): "ATF3 increases the intrinsic growth state of DRG neurons to enhance peripheral nerve regeneration," *J. Neurosci.*, 27, 7911–7920, URL http://dx.doi.org/10.1523/JNEUROSCI.5313-06.2007.

Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher (1998): "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." *Mol Biol Cell*, 9, 3273–97, URL http://www.molbiolcell.org/cgi/content/full/9/12/3273.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008): "Conditional variable importance for random forests," *BMC Bioinformatics*, 9, 307, URL http://www.biomedcentral.com/1471-2105/9/307.

Sugiura, N. (1978): "Further analysis of the data by akaike's information criterion and the finite corrections." *Communications in Statistics - Theory and Methods*, 7, 13–26, URL http://dx.doi.org/10.1080/03610927808827599.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman (2001): "Missing value estimation methods for dna microarrays." *Bioinformatics*, 17, 520–525, URL http://dx.doi.org/10.1093/bioinformatics/17.6.520.

Tsai, H.-K., H. H.-S. Lu, and W.-H. Li (2005): "Statistical methods for identifying yeast cell cycle transcription factors," *Proceedings of the National Academy of Sciences of the United States of America*, 102, 13532–13537, URL http://www.pnas.org/content/102/38/13532.abstract.

Van der Laan, M. (2006): "Statistical inference for variable importance," *The International Journal of Biostatistics*, 2, URL http://ideas.repec.org/a/bpj/ijbist/v2y2006i1n2.html.

Van der Laan, M. and D. Rubin (2006): "Targeted maximum likelihood learning," *International Journal of Biostatistics*, 2, 1043–1043, URL http://ideas.repec.org/a/bep/ijbist/v2y2006i1p1043-1043.html.

Wang, L., J. Zhu, and H. Zou (2006): "The doubly regularized support vector machine," *Statistica Sinica*, 16, 589–615.

Wu, W.-S. and W.-H. Li (2008): "Systematic identification of yeast cell cycle transcription factors using multiple data sources," *BMC Bioinformatics*, 9, 522, URL http://www.biomedcentral.com/1471-2105/9/522.

Zhang, N. R., M. C. Wildermuth, and T. P. Speed (2008): "Transcription factor binding site prediction with multivariate gene expression data," *The Annals of Applied Statistics*, 2, 332–365.

Zou, H. and T. Hastie (2005): "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society B*, 67, 301–320, URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.89.1596.